Harvard Kennedy School Misinformation Review[1]
December 2021, Volume 2, Issue 6

DOI: https://doi.org/10.37016/mr-2020-85
Website: misinforeview.hks.harvard.edu

*Research Note*

# Research note: Examining how various social media platforms have responded to COVID-19 misinformation

*We analyzed community guidelines and official news releases and blog posts from 12 leading social media and messaging platforms (SMPs) to examine their responses to COVID-19 misinformation. While the majority of platforms stated that they prohibited COVID-19 misinformation, the responses of many platforms lacked clarity and transparency. Facebook, Instagram, YouTube, and Twitter had largely consistent responses, but other platforms varied with regard to types of content prohibited, criteria guiding responses, and remedies developed to address misinformation. Only Twitter and YouTube described their systems for applying various remedies. These differences highlight the need to establish general standards across platforms to address COVID-19 misinformation more cohesively.*

Authors: Nandita Krishnan (1), Jiayan Gu (1), Rebekah Tromble (2,3), Lorien C. Abroms (1,2)
Affiliations: (1) Department of Prevention and Community Health, Milken Institute School of Public Health, The George Washington University, USA, (2) Institute for Data, Democracy and Politics, The George Washington University, USA, (3) School of Media and Public Affairs, The George Washington University, USA
How to cite: Krishnan, N., Gu, J., Tromble, R., & Abroms, L. C. (2021). Research note: Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School (HKS) Misinformation Review*, *2*(6).
Received: September 17th, 2021. Accepted: December 2nd, 2021. Published: December 15th, 2021.

## Research questions

- Do SMPs prohibit COVID-19 misinformation? What types of COVID-19 content do they prohibit, and what criteria do they use to inform action on misleading content?
- What remedies have SMPs developed to address COVID-19 misinformation? How are different remedies applied?

## Essay summary

- We conducted a content analysis of community guidelines, official news releases, and blog posts published between February 1, 2020, and April 1, 2021, for 10 social media platforms (Facebook, YouTube, Twitter, Instagram, Reddit, Snapchat, LinkedIn, TikTok, Tumblr, and Twitch) and two messaging platforms (Messenger and WhatsApp). This initial analysis was updated by a rapid review of the same data sources on November 23, 2021.
- The majority of the SMPs explicitly prohibited COVID-19 misinformation ($N = 8$), but only four (Facebook, Instagram, YouTube, and Twitter) had an independent COVID-19 misinformation

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

policy. Twitch, Tumblr, Messenger, and WhatsApp did not prohibit COVID-19 misinformation.

- The majority of SMPs (*N* = 10) developed one or more remedies to address misinformation and connect users with credible information. Remedies included soft measures, such as attaching warning labels, and hard measures, such as content removal and account bans. Only YouTube and Twitter described their systems for applying these different remedies, with progressively harsher remedies applied to repeat violators.
- The lack of clarity and transparency from many SMPs regarding their responses to COVID-19 misinformation and systems for applying various remedies makes it difficult for policymakers, researchers, and the general public to determine whether platforms are doing enough to address COVID-19 misinformation.
- Establishing a general set of policies and practices might be necessary to address COVID-19 misinformation in the broader social media ecosystem.

# Implications

Efforts to contain the COVID-19 pandemic have been complicated by a parallel infodemic, defined as an overabundance of information, including misinformation, which occurs during a disease outbreak (WHO, 2021). With over 3.8 billion users globally (DataReportal, 2021), social media and messaging platforms (hereafter referred to as SMPs) have become one of the major means of seeking and sharing COVID-19–related information. While the wide reach of SMPs has benefits in democratizing information access, it has also contributed to facilitating the rapid spread of mis- and disinformation (Cinelli et al., 2020; Kouzy et al., 2020). Some examples of COVID-19 misinformation[2] that have spread widely on SMPs include claims that 5G causes the virus and that COVID-19 vaccines alter DNA (Islam et al., 2021, 2020; Naeem et al., 2021). Such types of misinformation have been widely viewed and shared on SMPs (Nielsen et al., 2020). A recent study found that exposure to such types of misinformation was negatively associated with vaccination intention (Loomba et al., 2021), illustrating that misinformation has real consequences for containing the pandemic.

The magnitude of the COVID-19 pandemic has resulted in urgent calls from all sectors of society for SMPs to do more to address COVID-19 misinformation (*Disinformation nation: Social media's role in promoting extremism and misinformation*, 2021; Donovan, 2020). However, SMPs face a number of challenges in determining how best to do this, including the massive volume of content and nuanced nature of misinformation. Thus, SMPs will need to decide what types of content to prioritize, when to take action, and what type of action to take. Legal scholars have proposed typologies of content moderation remedies in internet platforms (Goldman, 2021; Grimmelmann, 2015). These include more lenient "soft" remedies (e.g., warning labels) and more stringent "hard" remedies (e.g., removing content) (Goldman, 2021; Grimmelmann, 2015). Given that the different types of remedies have different consequences for users, it is important to understand how platforms apply these remedies. Additionally, as many SMP users use multiple platforms (DataReportal, 2021), as newer platforms are gaining popularity, and as platforms are increasingly interconnected (e.g., TikTok videos can be shared on Facebook), effectively addressing misinformation requires responses from all platforms.

We examined responses to COVID-19 misinformation by 12 leading SMPs and found that the majority (eight out of 12) prohibited COVID-19 misinformation. This finding by itself is noteworthy, as for

[2] In the interest of brevity, in this paper we use the term "misinformation" broadly to cover information and claims that are both intentionally and unintentionally false and misleading. As noted elsewhere in this paper, the platforms we examine define the scope of "misinformation" in various ways.

several SMPs, it represents a stark reversal from their previous stance on content regulation. As recently as two years ago, companies such as Facebook[3] and Twitter refused to act on misinformation, arguing that doing so would infringe on free speech and was outside of their platform mission (Conger, 2019; Kang & Isaac, 2019). While this change is encouraging, there is much room for improvement in how SMPs address COVID-19 misinformation.

Descriptions of the responses of many SMPs lacked clarity, and their implementation lacked transparency. Many SMPs did not clearly articulate the types of content prohibited. This might be partly attributable to the challenges inherent in moderating scientific content, which is dynamic and often equivocal (Baker et al., 2020). In the COVID-19 context, this has been reflected in evolving scientific positions regarding the nature of COVID-19 transmission (WHO, 2020) and lack of consensus on boosters (Krause et al., 2021), to give just a few examples. Challenges notwithstanding, clearly outlining prohibited content is important. It can help researchers assess whether the types of content SMPs prohibit correspond with the types of content associated with real-world harm. Such evidence could guide SMPs on prioritizing and updating the types of content they prohibit. For instance, Facebook initially did not prohibit personal anecdotes or first-person accounts, but evidence showed that this type of content might be contributing to vaccine hesitancy (Dwoskin, 2021). Facebook has since updated its policy on content related to COVID-19 vaccines and now reduces the distribution of alarmist or sensationalist content about vaccines. Additionally, few platforms have a specific COVID-19 misinformation policy. During an evolving emergency such as a pandemic, such a policy could make it possible for platforms to communicate their stance on various types of content, changes to their responses, and the consequences of violating policies in a transparent and accessible manner.

Similar to previous studies (Nunziato, 2020; Sanderson et al., 2021), we found that most SMPs employed a combination of soft and hard remedies to address misinformation. Most platforms, however, did not clearly describe their systems for applying these different types of remedies. Clear visibility of consequences can deter bad behavior (Seering et al., 2017), and the U.S. surgeon general's report on health misinformation calls for platforms to "impose clear consequences for accounts that repeatedly violate platform policies" (Office of the Surgeon General, 2021). YouTube's and Twitter's strike systems are consistent with the responsive regulation approach, which advocates using soft remedies before escalating to hard remedies (Ayres & Braithwaite, 1992). Given the findings of a recent report that 12 people were responsible for 73% of misinformation on SMPs (Center for Countering Digital Hate, 2021; Dwoskin, 2021), this approach could be well suited to deter superspreaders and repeat violators. Interestingly, Facebook refuted the conclusion of that report, claiming that these 12 individuals were responsible for only 0.05% of all views of vaccine-related content on that platform (Bickert, 2021). Additional analyses by independent researchers could provide more evidence regarding the extent to which specific individuals contribute to the spread of misinformation and the effectiveness of responsive regulation in curbing the spread of misinformation. More transparency from platforms regarding their application of different remedies is key to determining whether they are doing enough to address misinformation.

Facebook, Instagram, YouTube, and Twitter had largely similar responses, but the responses of other platforms differed. There is recognition among scholars that a one-size-fits-all approach to content moderation is not practical as SMPs vary with regard to their functions, audiences, and capacity to moderate content (Gillespie et al., 2020; Goldman, 2021). For instance, the encrypted nature of communication on messaging platforms means that content moderation approaches for these platforms will need to preserve user privacy (Gillespie et al., 2020). Thus, total alignment of responses across

[3] Facebook, Inc. changed its name to Meta on October 28, 2021. As the company's name was Facebook at the time of the initial analysis, we have used the name Facebook, when referring to Facebook, Inc.

platforms is not feasible. However, previous research has shown that even if some platforms act on misinformation, it can continue to spread on other platforms (Sanderson et al., 2021). Additionally, conspiracy theorists might be migrating to alternative platforms (e.g., Parler, Twitch) as a result of mainstream platforms (e.g., Facebook, YouTube) cracking down on misinformation (Browning, 2021). Therefore, some degree of synergy in platform responses might be necessary to address COVID-19 misinformation in the broader social media ecosystem. Public–private co-regulation models are considered well suited to achieve this (Gillespie, 2018; Gorwa, 2019). Platforms could collaborate for mutual benefit and have taken steps in that direction (Shu & Shieber, 2020). Public policies could outline general standards that apply to all SMPs.

While the policies and remedies identified by this analysis represent an important first step, enforcement is vital for them to have an impact. Some platforms, such as Facebook, Instagram, YouTube, and Twitter, release some enforcement metrics in their transparency reports, such as the number of posts taken down or accounts disabled. However, without a meaningful denominator, it is impossible to determine the extent to which these policies are being enforced. External analyses indicate poor enforcement of misinformation policies by SMPs. For instance, one analysis found that a large proportion of content rated false by fact-checkers remained up on Twitter (59%), and a substantial amount also remained up on YouTube (27%) and Facebook (24%) (Brennen et al., 2020). A number of analyses have examined enforcement of misinformation policies by Facebook and identified several gaps. One report found that only 16% of all health information analyzed on Facebook had a warning label (Avaaz, 2020a). Even when Facebook took action, it was slow in doing so, taking up to 22 days to downgrade or attach warning labels to misleading content (Avaaz, 2020b). There are also significant language gaps in enforcement of policies. One analysis found that approximately 70% of misleading COVID-19 content in Spanish and Italian (vs. 29% of misleading content in English) had not received warning labels on Facebook (Avaaz, 2020b). Enforcement also varies by type of content. Although Facebook's vaccine misinformation policy applies to all content on the platform, warning labels are not applied to content in the comments section. A case study found that misinformation is rife in the Facebook comments section, with roughly one in five comments found to contain misinformation about the vaccines or the pandemic (Chan et al., 2021).

Additionally, recent revelations in the Facebook Papers, such as special treatment given by Facebook to high-profile users when it comes to content moderation (Allyn et al., 2021), hinder the credibility of any enforcement claims made by SMPs. There are growing calls for platforms to make their responses more transparent and data more easily available to researchers (Doctors for America, 2021; MacCarthy, 2020; Office of the Surgeon General, 2021). Such independent monitoring of enforcement of policies and remedies, and evaluation of specific remedies (e.g., labeling content) and their systems of application (e.g., strike system) identified by this analysis, could hold SMPs accountable. By laying out COVID-19 misinformation responses and remedies used by a wide range of SMPs, this work can also serve as a starting point to help policymakers identify a general set of policies that might be applicable to all platforms, as well as those that might be applicable to similar types of platforms (e.g., messaging platforms).

## Evidence

Table 1 summarizes SMP responses to COVID-19 misinformation (see Appendix Table 1 for examples). The majority of SMPs explicitly prohibited COVID-19 misinformation ($N = 8$), but only four (Facebook, Instagram, YouTube, and Twitter) had an independent COVID-19 misinformation policy. Twitch, Tumblr, Messenger, and WhatsApp did not prohibit COVID-19 misinformation. Half the SMPs delineated one or more types of COVID-19 content that they prohibited (Table 1, Appendix Table 2). Facebook, Instagram,

YouTube, and Twitter all prohibited misinformation regarding the nature of the virus, the efficacy and safety of prevention and treatment measures, COVID-19 vaccines, and restrictions and health advisories, as well as content that misrepresented data. LinkedIn prohibited content that fell under the first three of the above categories. TikTok did not publish a detailed list of types of COVID-19 misinformation it prohibited, but it stated that COVID-19 vaccine misinformation was prohibited. Snapchat and Reddit did not articulate the specific types of COVID-19 content they prohibited.

***Table 1.*** *Summary of responses to COVID-19 misinformation by leading social media and messaging platforms.*

| | **Face-book** | **Insta-gram** | **You-Tube** | **Twit-ter** | **Linked-In** | **Snap-chat** | **Red-dit** | **Tik-Tok** | **Twitch** | **Tumblr** | **Mes-senger** | **Whats-App** | **Total (*N* = 12)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Response to COVID-19 misinformation*** | | | | | | | | | | | | | |
| Pro-hibits COVID-19 mis-infor-mation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | 8 |
| Has a COVID-19 mis-infor-mation policy | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | × | 4 |
| ***Types of COVID-19 misinformation prohibited*** | | | | | | | | | | | | | |
| Nature of the virus | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | 5 |
| Efficacy & sa-fety of preven-tion & treat-ment mea-sures | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | 5 |
| Vaccine mis-infor-mation | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | ✓ | × | × | × | × | 6 |
| Restric-tions & health advi-sories | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | × | 4 |
| Mis-repre-sen-tation of data | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | × | 4 |
| ***Criteria used to inform action on COVID-19–related content*** | | | | | | | | | | | | | |
| Claim de-bunked by | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | 5 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| public health authorities | | | | | | | | | | | | | |
| Potential of the content to lead to significant harm | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | × | × | × | × | 7 |
| Authenticity | ✓[1] | ✓[1] | ✓ | ✓ | × | ✓[1*] | ✓ | ✓ | × | × | × | ✓ | 8 |
| Other | ✓[2] | ✓[2] | ✓[3] | ✓[4] | × | × | × | × | × | × | × | × | 4 |

*Notes: ✓ indicates presence and × indicates absence of response. [1]These platforms have policies prohibiting coordinated inauthentic behavior as part of their broader community guidelines and terms of use. These policies apply to all types of content, but are not explicitly mentioned in connection with COVID-19 content. [2]Facebook and Instagram's COVID-19 policies also apply to content related to the "coordination of harm, sale of medical masks and related goods, hate speech, bullying and harassment". [3]YouTube also acts on borderline content—"content that comes close to, but doesn't quite cross the line of violating" its Community Guidelines—that could misinform users in harmful ways. [4]Twitter also acts on misleading content that is expressed as an assertion of fact or incites people to action that could result in physical/social harm or damage to infrastructure. [*]This criterion was identified in the updated rapid review conducted on November 23, 2021.*

Most platforms (*N* = 9) listed at least one criterion they used to guide action on COVID-19–related content (Table 1, Appendix Table 3). A majority of SMPs indicated that they would take action on content that had the potential to cause significant harm (*N* = 7) or was inauthentic (*N* = 8). Five SMPs also acted on COVID-19 content if it contained claims that had been debunked by public health authorities. Some platforms outlined other criteria. For instance, Facebook and Instagram prohibited content that was considered to be hate speech or harassment, YouTube also acted on borderline content, (i.e., content that almost but not actually violated their misinformation policy), and Twitter took action against misleading COVID-19 content that could cause social unrest or was expressed as an assertion of fact.

The majority of SMPs developed one or more remedies to address misinformation (*N* = 10) and proactively connect users with credible information (*N* = 10). Table 2 summarizes the former type of remedies (see Appendix Table 4 for examples). Seven SMPs used labels, warnings, notifications, and links. A different group of seven SMPs imposed restrictions on advertisements that had the potential to spread misinformation or cause harm. Eight SMPs modified their search and recommendation algorithms or disabled "like," "share," and forwarding features to reduce the visibility and distribution of misleading content. However, platforms differed slightly in how they used this remedy. For instance, Facebook and Instagram stated that they would reduce the distribution of content from users that had repeatedly shared misleading content or previously violated their policies, while YouTube also applied this remedy to borderline content. Eight SMPs removed some types of content, and seven SMPs temporarily suspended or permanently banned accounts or groups.

***Table 2.*** *Remedies to address COVID-19 misinformation across leading social media and messaging platforms.*

| | **Face-book** | **Insta-gram** | **You-Tube** | **Twit-ter** | **Linked-In** | **Snap-chat** | **Red-dit** | **Tik-Tok** | **Twitch** | **Tumblr** | **Mes-senger** | **Whats-App** | **To-tal (*N* = 12)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Labels, warnings, notifycations, & links | ✓ | ✓ | × | ✓ | × | × | ✓ | ✓ | × | × | ✓[1] | ✓[1] | 7 |
| Decreasing visibility & spread | ✓ | ✓ | ✓ | ✓ | × | × | ✓ | ✓ | × | × | ✓[2] | ✓[2] | 8 |
| Content removal | ✓[3] | ✓ | ✓ | ✓ | ✓ | ✓* | ✓[3] | ✓ | × | × | × | × | 8 |
| Account suspension/ban | ✓ | ✓ | ✓ | ✓ | × | × | ✓* | ✓[4] | × | × | ×[5] | ✓ | 7 |
| Ad restrictions | ✓ | ✓ | ✓ | ✓ | ✓ | ✓* | × | ✓ | × | × | N/A | N/A | 7 |

*Notes: ✓ indicates presence and × indicates absence of remedy; N/A indicates remedy not applicable. [1]Messenger and WhatsApp apply forwarding labels to indicate to recipients that the message was not written by the sender; WhatsApp also allows users to fact check the content of viral messages through Google search. [2]Messenger and WhatsApp apply forwarding limits, which limit the number of people or groups a message can be forwarded to at one time. [3]Misinformation in Facebook groups and Reddit communities can also be removed by moderators. [4]Applies to users who violate any of TikTok's Community Guidelines and Terms of Service, including their misinformation policies. [5]Because of how Messenger is connected to Facebook, Facebook account bans may apply to Messenger; however, this was not explicitly mentioned. *This remedy was identified in the updated rapid review conducted on November 23, 2021.*

Most SMPs did not clearly describe the manner in which they implemented these wide-ranging remedies, except for YouTube and Twitter, which used a strike system. Under this system, individuals received strikes for each violation. For instance, YouTube provided a warning to first-time violators with no strikes. However, if a user had previously violated the COVID-19 content policy, each subsequent violation resulted in a strike, with three strikes leading to channel termination. On Twitter, content that was labeled received one strike, while content that was deleted received two strikes. For the first strike, no account-level action was taken. For each subsequent strike, the following actions were taken: 12-hour account lock (two or three strikes), seven-day account lock (four strikes), and permanent suspension (five or more strikes). Reddit did not have a formal strike system but implied that it also took action in a graduated manner by stating that "our goal is always to start with education and cooperation and only escalate to quarantine or ban if necessary." Similarly, Facebook stated that profiles "that repeatedly post misinformation related to COVID-19, vaccines, and health may face restrictions, including (but not limited to) reduced distribution, removal from recommendations, or removal from our site," but it was unclear on what basis these different remedies were applied.

Actions to proactively connect users with credible information are summarized in Table 3 (see Appendix Table 5 for examples). Common measures included information curation (*N* = 9); labels, banners and links (*N* = 8); and Q&As with experts (*N* = 7). Six SMPs used health promotion campaigns to increase awareness and promote appropriate practices advised by health authorities, such as mask-wearing, handwashing, and social distancing. The nature of campaigns tended to vary by platform design. For instance, platforms that allowed users to post temporary stories, such as Instagram and Snapchat, launched stickers and filters to help people share recommended COVID-19 prevention

practices and encourage vaccination. Video-based platforms, such as YouTube, launched video campaigns in collaboration with external public health partners to reach underserved populations. Meanwhile, Twitter promoted hashtags, such as #vaccinated and #WearAMask, and TikTok used hashtags to promote challenges, such as the #safehandschallenge. Facebook, Instagram, YouTube, Twitter, and LinkedIn described adjustments to their algorithms to increase visibility of authoritative content in searches and recommendations. These five platforms and TikTok also provided free advertising credits to public health partners for disseminating COVID-19 information.

***Table 3.*** *Remedies to promote access to evidence-based COVID-19 information across leading social media and messaging platforms.*

| | **Face-book** | **Insta-gram** | **You-Tube** | **Twit-ter** | **Linked-In** | **Snap-chat** | **Red-dit** | **Tik-Tok** | **Twitch** | **Tumblr** | **Mes-senger** | **Whats App** | **Total (*N* = 12)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Information curation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | ✓ | × | × | ✓[1] | ✓[1] | 9 |
| Health promotion campaigns | ✓ | ✓ | ✓ | ✓ | × | ✓ | × | ✓ | × | × | × | × | 6 |
| Labels, banners & links | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | 8 |
| Increasing visibility of authoritative content | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | × | × | × | × | × | 5 |
| Q&As with experts | ✓ | × | ✓ | ✓ | ✓ | ✓[2] | ✓ | ✓ | × | × | × | × | 7 |
| Ad credits to public health partners | ✓ | ✓ | ✓ | ✓ | ✓ | × | × | ✓ | × | × | × | × | 6 |

*Notes: ✓ indicates presence and* **×** *indicates absence of remedy. [1]Messenger and WhatsApp have messaging helplines that allow users to get information directly from the WHO. [2]It was not clear whether Q&As were live or pre-recorded.*

Due to the end-to-end encrypted features and manner in which information is shared on messaging platforms, these platforms had different strategies to deal with misinformation compared to traditional social media platforms. While Messenger and WhatsApp did not explicitly state that they prohibited COVID-19 misinformation, they implemented some actions to limit its spread. These included attaching forwarding labels to messages that did not originate with the sender and introducing forwarding limits to reduce the spread of viral messages. Additionally, both platforms collaborated with the WHO to provide users with accurate and timely information about COVID-19 via free messaging. WhatsApp also collaborated with external fact-checking organizations and used advanced machine learning approaches to identify and ban accounts engaged in mass messaging. It was unclear whether Facebook account bans also applied to Messenger.

Given the evolving SMP policy environment, we undertook a rapid review of community guidelines,

news releases, and blog posts for these 12 SMPs on November 23, 2021, to determine whether there were any changes to their responses to COVID-19 misinformation since our initial analysis. While the majority of SMPs did not have major updates to their responses, a few platforms had some noteworthy changes. Facebook had begun to reduce the distribution of misleading vaccine-related content that did not violate its policies but had the potential to discourage vaccinations. Facebook also stated that it would target certain remedies at misinformation superspreaders. These primarily included soft remedies, such as providing a label or warning when users liked a page found to repeatedly share misinformation, decreasing visibility of posts in the News Feed from individual Facebook accounts repeatedly spreading misinformation, and notifying users that their posts might be demoted in the News Feed if they repeatedly shared misinformation. Facebook also applied some hard remedies, such as removal of some Facebook pages and groups and Instagram accounts linked to 12 high-volume superspreaders identified in an external report (Center for Countering Digital Hate, 2021), despite questioning the findings of the report.

Other platforms that have updated or provided more transparency about their responses to addressing COVID-19 misinformation since our initial analysis are YouTube, Snapchat, Twitter, and Reddit. YouTube extended its ban on COVID-19 vaccine misinformation to misinformation about all vaccines that are approved and considered safe and effective by the WHO and other health authorities. In August 2021, Snapchat released more information regarding its approach to handling misinformation, which differs from other platforms' approaches in a few ways. For example, Snapchat's newsfeed is proactively moderated; group chats are not recommended by algorithms; and, rather than applying soft remedies, all content that violates guidelines is removed. Twitter announced that it was testing a new feature that allowed users to report content as misleading. Additionally, Twitter began collaborating with the Associated Press and Reuters to provide more context to conversations and improve information curation. Reddit reversed its decision to not act on misinformation by banning a subreddit. However, the subreddit was banned on the grounds of harassment rather than misinformation, and it remains to be seen whether Reddit takes further steps to address misinformation.

# Methods

*Data*

We conducted a content analysis of documents from 10 leading social media platforms and two messaging platforms in the United States (U.S.). There is no single database that monitors the volume of users across all SMPs, making it difficult to rank platforms in terms of popularity. However, a recent nationally representative survey assessed prevalence of use of 10 social media platforms (Facebook, YouTube, Twitter, Instagram, Reddit, Snapchat, LinkedIn, TikTok, Tumblr, and Twitch), and one messaging platform (WhatsApp) (Shearer & Mitchell, 2021). Prevalence of use ranged from 74% for YouTube to 4% for Tumblr (Shearer & Mitchell, 2021). This group of SMPs[4] was chosen because prevalence data was available.

Data sources included: (i) terms of use and community guidelines and (ii) blog posts and news releases published on the official website of each SMP between February 1, 2020, and April 1, 2021. All SMP websites provided a listing of blog posts and news releases in reverse chronological order. All blog posts and news releases published within this time frame were reviewed, and those that were relevant

[4] We analyzed Messenger independently from Facebook, as Messenger is a messaging platform, and approaches to addressing misinformation by messaging platforms will likely differ from approaches used by social media platforms. Hence, our sample consisted of 12 SMPs.

to addressing COVID-19 misinformation or misinformation, in general, were retained for further analysis. To be included, documents had to contain descriptions of any of the following: (1) the platform's response to COVID-19 misinformation or misinformation generally, (2) types of COVID misinformation it prohibited and criteria it used to make this determination, or (3) remedies developed to address misinformation or proactively connect users to credible information. We included responses and remedies that were explicitly linked to addressing COVID-19 misinformation as well as those intended to address misinformation broadly, under the assumption that broader misinformation responses and remedies also applied to COVID-19 misinformation. One analyst conducted a rapid review using the same data sources for all platforms on November 23, 2021, to determine whether any SMPs had updated their responses to COVID-19 misinformation.

*Development of the codebook*

As there are no well-established frameworks for misinformation classification and intervention in the digital environment, a codebook was developed using an inductive approach. One analyst reviewed documents from a single platform (Twitter) and developed an initial codebook.

*Description of codes*

- *Response to COVID-19 misinformation:* Codes under this category included whether the platform prohibited COVID-19 misinformation and whether it had a COVID-19 misinformation policy.
- *Types of COVID-19 misinformation prohibited:* included false or misleading content about (i) the nature of the virus (including the existence, origin, causes, diagnosis, and transmission of COVID-19), (ii) the efficacy and safety of prevention and treatment measures (e.g., hydroxychloroquine), (iii) COVID-19 vaccines, (iv) restrictions and health advisories (including established mitigation measures such as masks, social distancing and handwashing), and (v) misrepresentation of data (e.g., prevalence of COVID-19 in an area or availability of resources such as hospital beds).
- *Criteria used to inform action on COVID-19 related content:* included (i) debunking of claims by public health authorities, (ii) potential of the content to lead to significant harm, (iii) authenticity of content (e.g., platform manipulation, fake accounts, or deep fakes), and (iv) other.
- *Actions to address misinformation:* included (i) labels, warnings, and notifications to inform users that the content they are viewing or sharing is misleading, and links to credible organizations such as the World Health Organization (WHO); (ii) decreasing visibility and spread of misleading content by lowering such content in searches and newsfeeds, and by restricting a user's ability to engage with or share such content; (iii) content removal; (iv) temporary account bans or permanent account suspensions; and (v) advertising restrictions to limit the promotion of unverified cures or products with unverified claims regarding their prevention and treatment efficacy.
- *Actions to promote access to credible information:* included (i) information curation, whereby SMPs curate and compile credible information that is easily accessible to users (e.g., a resource center or newsletter); (ii) health promotion and communication campaigns to promote evidence-based measures to contain the pandemic, such as vaccination, social distancing, and mask-wearing; (iii) labels, banners and links to credible organizations, such as WHO or curated information hubs within the platform proactively provided to users; (iv) increasing visibility of authoritative content by elevating such content in searches and newsfeeds; (v) Q&As with public health experts through chats and live streams; and (vi) advertising credits to government and public health organizations to disseminate COVID-19 information.

*Coding and analysis*

For each SMP, each of the codes described above was marked as present if it was described in their documents at least once, and absent if it was not mentioned. As we were interested in identifying all types of responses used by SMPs to address COVID-19 misinformation, we coded a policy or remedy as present if it was in effect at any point during our time frame of interest (February 1, 2020, to April 1, 2021), even if it was subsequently removed. It should be noted that Instagram, Messenger, and WhatsApp are owned by Facebook, and while some Facebook policies apply to all these platforms, it is not clear which do. Therefore, we treated all these platforms as separate entities. When coding for Instagram, Messenger, and WhatsApp, we applied text segments from Facebook's documents to these platforms only if the text explicitly referenced the platforms. Coding and analysis occurred in an iterative fashion. Two analysts independently coded documents from six SMPs using the initial version of the codebook, met to discuss their findings, and modified the codebook. Both analysts then independently coded documents from all 12 SMPs using the modified codebook. Cohen's kappa was calculated to assess intercoder reliability for coding pertaining to each table and for all tables combined, with raters achieving high agreement (overall $\kappa$: 0.94, 95% CI: 0.90–0.99; Table 1 $\kappa$: 0.95, 95% CI: 0.89–1.00;[5] Table 2 $\kappa$: 0.96, 95% CI: 0.88–1.00; Table 3 $\kappa$: 0.93, 95% CI: 0.83–1.00). The analysts met again to discuss their coding, made minor changes to the codebook, and resolved all disagreements through discussion. The presence of each code was summed across platforms to facilitate comparisons.

*Limitations*

A few limitations must be noted. We did not analyze news articles, which might have captured additional SMP responses and remedies not noted on their websites. However, media coverage could be uneven and biased towards bigger platforms, such as Facebook and Twitter. We, therefore, decided to limit our analysis to documents published by SMPs. Some of these policies were dynamic and changed over the course of the pandemic, but our analysis was not designed to capture such changes over time. Lack of clarity from Facebook regarding the relationship between Facebook's Community Standards and Community Guidelines and its other platforms, such as Instagram, may have resulted in miscoding certain responses and remedies as absent when they were actually present on these platforms. It is possible that our search strategy could have missed certain documents, which could also have resulted in miscoding certain responses or remedies as absent when they might have been present. Finally, the descriptive nature of this analysis precludes us from making any conclusions about the effectiveness or implementation of policies for any platform, but these findings can inform future work on these topics.

# Bibliography

Allyn, B. (2021, October 21). *Oversight board slams Facebook for giving special treatment to high-profile users.* NPR. https://www.npr.org/2021/10/21/1047928585/oversight-board-slams-facebook-for-giving-special-treatment-to-vip-users

Avaaz. (2020a). *Facebook's algorithm: A major threat to public health.* https://secure.avaaz.org/campaign/en/facebook_threat_health

Avaaz. (2020b). *How Facebook can flatten the curve of the coronavirus infodemic.* https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation

[5] For Table 1, the code labeled "other" was not included in the calculation of $\kappa$, as this code was added to the codebook and defined through discussion between the two analysts after the second round of coding.

Ayres, I., & Braithwaite, J. (1992). *Responsive regulation: Transcending the deregulation debate*. Oxford University Press.

Baker, S. A., Wade, M., & Walsh, M. J. (2020). The challenges of responding to misinformation during a pandemic: Content moderation and the limitations of the concept of harm. *Media International Australia*, *177*(1), 103–107. https://doi.org/10.1177/1329878X20951301

Bickert, M. (2021, August 18). How we're taking action against vaccine misinformation superspreaders. https://about.fb.com/news/2021/08/taking-action-against-vaccine-misinformation-superspreaders

Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation.* Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Browning, K. (2021, April 27). Extremists find a financial lifeline on Twitch. *The New York Times.* https://www.nytimes.com/2021/04/27/technology/twitch-livestream-extremists.html

Center for Countering Digital Hate. (2021, March 24). *The disinformation dozen*. https://www.counterhate.com/disinformationdozen

Chan, E., Beaman, L., & Zhang, S. (2021, May 6). *Vaccine misinformation in Facebook comment sections: A case study.* First Draft. https://firstdraftnews.org/articles/vaccine-misinformation-in-facebook-comment-sections-a-case-study

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Conger, K. (2019, November 2). Twitter stands by Trump amid calls to terminate his account. *The New York Times.* https://www.nytimes.com/2019/10/15/technology/trump-twitter-account.html

DataReportal. (2021). *Global social media stats*. Retrieved August 3, 2021, from https://datareportal.com/social-media-users

*Disinformation nation: Social media's role in promoting extremism and misinformation: Hearing before the U.S. House Committee on Energy and Commerce*, 117th Cong. (2021). https://www.congress.gov/event/117th-congress/house-event/111407

Doctors for America. (2021, November 23). *Letter to Facebook: Disclose your data now*. https://doctorsforamerica.org/letter-to-facebook-disclose-your-data-now

Donovan, J. (2020, April 14). Social-media companies must flatten the curve of misinformation. *Nature*. https://doi.org/10.1038/d41586-020-01107-z

Dwoskin, E. (2021, March 14). Massive Facebook study on users' doubt in vaccines finds a small group appears to play a big role in pushing the skepticism. *The Washington Post.* https://www.washingtonpost.com/technology/2021/03/14/facebook-vaccine-hesistancy-qanon

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S., Sinnreich, A., & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1512

Goldman, E. (2021). *Content moderation remedies*. SSRN. http://dx.doi.org/10.2139/ssrn.3810580

Gorwa, R. (2019, October 28). Regulating them softly. In *Models for platform governance*, 39–43. Centre for International Governance Innovation. https://www.cigionline.org/models-platform-governance/

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, *17*(1), 42–108. https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1110&context=yjolt

Islam, M. S., Mostofa Kamal, A.-H., Kabir, A., Southern, D. L., Khan, S. H., Hasan, S. M. M., Sarkar, T., Sharmin, S., Das, S., Roy, T., Harun, M. G. D., Chughtai, A. A., Homaira, N., & Seale, H. (2021). COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence. *PloS ONE*, *16*(5), e0251605. https://doi.org/10.1371/journal.pone.0251605

Islam, M. S., Sarkar, T., Khan, S. H., Mostofa Kamal, A.-H., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., & Seale, H. (2020). COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, *103*(4), 1621–1629. https://doi.org/10.4269/ajtmh.20-0812

Kang, C., & Isaac, M. (2019, October 21). Defiant Zuckerberg says Facebook won't police political speech. *The New York Times.* https://www.nytimes.com/2019/10/17/business/zuckerberg-facebook-free-speech.html

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, *12*(3), e7255. https://doi.org/10.7759/cureus.7255

Krause, P. R., Fleming, T. R., Peto, R., Longini, I. M., Figueroa, J. P., Sterne, J. A. C., Cravioto, A., Rees, H., Higgins, J. P. T., Boutron, I., Pan, H., Gruber, M. F., Arora, N., Kazi, F., Gaspar, R., Swaminathan, S., Ryan, M. J., & Henao-Restrepo, A.-M. (2021). Considerations in boosting COVID-19 vaccine immune responses. *The Lancet*, *398*(10308), 1377–1380. https://doi.org/10.1016/S0140-6736(21)02046-8

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

MacCarthy, M. (2020). *Transparency requirements for digital social media platforms: Recommendations for policy makers and industry.* SSRN. http://dx.doi.org/10.2139/ssrn.3615726

Naeem, S. B., Bhatti, R., & Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information and Libraries Journal*, *38*(2), 143–149. https://doi.org/10.1111/hir.12320

Nielsen, R. K., Fletcher, R., Newman, N., Brennen, J. S., & Howard, P. N. (2020, April 15). *Navigating the 'infodemic': How people in six countries access and rate news and information about coronavirus.* Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/infodemic-how-people-six-countries-access-and-rate-news-and-information-about-coronavirus

Nunziato, D. C. (2020). Misinformation mayhem: Social media platforms' efforts to combat medical and political misinformation. *First Amendment Law Review*, *19*, 32. https://scholarship.law.gwu.edu/faculty_publications/1502/

Office of the Surgeon General. (2021). *Confronting health misinformation: The U.S. Surgeon General's advisory on building a healthy information environment*. U.S. Department of Health and Human Services. https://www.surgeongeneral.gov/healthmisinformation

Sanderson, Z., Brown, M. A., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School (HKS) Misinformation Review*, *2*(4). https://doi.org/10.37016/mr-2020-77

Seering, J., Kraut, R., & Dabbish, L. (2017). Shaping pro and anti-social behavior on Twitch through moderation and example-setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing,* pp. 111–125. https://dl.acm.org/doi/10.1145/2998181.2998277

Shearer, E., & Mitchell, A. (2021, January 12). *News use across social media platforms in 2020.* Pew Research Center. https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/

Shu, C., & Shieber, J. (2020, March 16). *Facebook, Reddit, Google, LinkedIn, Microsoft, Twitter and YouTube issue joint statement on misinformation.* Tech Crunch. https://techcrunch.com/2020/03/16/facebook-reddit-google-linkedin-microsoft-twitter-and-youtube-issue-joint-statement-on-misinformation/

World Health Organization. (2020). *Transmission of SARS-CoV-2: Implications for infection prevention precautions.* https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions

World Health Organization. (2021). *Infodemic.* https://www.who.int/health-topics/infodemic

**Funding**
This research is supported by the John S. and James L. Knight Foundation through a grant to the Institute for Data, Democracy & Politics at The George Washington University.

**Competing interests**
Rebekah Tromble has received funding from Facebook and Twitter in support of her research. All other authors have no competing interests.

**Ethics**
As this study consisted of analysis of publicly available documents and did not meet the definition of human subjects research, IRB approval was not required.



**Data availability**
All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/GCOJDX

# Appendix: Coded excerpts illustrating social media platform responses to COVID-19 misinformation

***Table 1.*** *Coded excerpts of responses to COVID-19 misinformation by leading social media and messaging platforms.*

| | ***Prohibits COVID-19 misinformation*** | ***COVID-19 misinformation policy*** |
|---|---|---|
| Facebook, Instagram | As people around the world confront this unprecedented public health emergency, we want to make sure that our Community Guidelines protect people from harmful content and new types of abuse related to COVID-19. We're working to remove content that has the potential to contribute to real-world harm, including through our policies prohibiting coordination of harm, sale of medical masks and related goods, hate speech, bullying and harassment and misinformation that contributes to the risk of imminent violence or physical harm. As the situation evolves, we continue to look at content on the platform, assess speech trends, and engage with experts, and will provide additional policy guidance when appropriate to keep the members of our community safe during this crisis. | Facebook COVID-19 and Vaccine Policy Updates & Protections<br><br>Instagram COVID-19 and Vaccine Policy Updates and Protections |
| YouTube | YouTube doesn't allow content that spreads medical misinformation that contradicts local health authorities' (LHA) or the World Health Organization's (WHO) medical information about COVID-19. | YouTube COVID-19 medical misinformation policy |
| Twitter | You may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm. | Twitter COVID-19 misleading information policy |
| LinkedIn | We've always prohibited false and misleading content, but we recently updated our Professional Community Policies to be clear that information contradicting guidance from leading global health organizations and public health authorities is also not allowed on the platform. | N/A |
| Snapchat | Our guidelines prohibit Snapchatters and our partners from sharing content that deceives or deliberately spreads false information that causes harm, and we do not offer an open news feed where unvetted publishers or individuals have an opportunity to broadcast misinformation. | N/A |
| Reddit | Our site integrity team is using their existing tools and processes to investigate claims and signs of coordinated attempts to spread COVID-19 misinformation on Reddit. | N/A |
| TikTok | TikTok's Community Guidelines prohibit content that's false or misleading, including misinformation related to COVID-19 and vaccines and anti-vaccine disinformation more broadly. | N/A |

*Note: N/A indicates that the platform did not have a COVID-19 misinformation policy.*

***Table 2.*** *Coded excerpts of types of COVID-19 content prohibited by leading social media and messaging platforms.*

| | | |
|---|---|---|
| **Nature of the virus** | Facebook, Instagram | More specifically, we remove false information about:<br>The existence or severity of COVID-19. Acknowledging the existence and understanding the severity of COVID-19 is foundational to keeping people safe and aware of the dangers of this public health emergency. We remove claims that deny the existence of the disease or undermine the severity of COVID-19.<br>COVID-19 transmission and immunity: Understanding how COVID-19 is transmitted and who can be infected is a critical component of protecting people from getting or spreading the virus. Public health authorities state that COVID-19 can be transmitted in any location and primarily from person to person through small droplets from the nose or mouth, which are expelled when a person with COVID-19 coughs, sneezes or speaks. Public health authorities also agree that all people, regardless of age or other unique characteristics, can be infected with and spread COVID-19. We remove false claims about how and where COVID-19 can be transmitted and who can be infected. |
| | YouTube | Don't post content on YouTube if it includes any of the following:<br>Diagnostic misinformation: Content that promotes diagnostic methods that contradict local health authorities or WHO.<br>Transmission misinformation: Content that promotes transmission information that contradicts local health authorities or WHO.<br>Content that denies the existence of COVID-19. |
| | Twitter | We will label or remove false or misleading information about:<br>Transmission of the virus, such as false claims about asymptomatic spread, or false information about how it is transmitted indoors.<br>Susceptibility to the virus, for example claims that specific groups or people are more or less prone to be infected or to develop adverse symptoms on the basis of their nationality or religion.<br>Symptoms commonly associated with the virus, for example, misleading instructions on how to self-diagnose. |
| | LinkedIn | We've always prohibited false and misleading content, but we recently updated our Professional Community Policies to be clear that information contradicting guidance from leading global health organizations and public health authorities is also not allowed on the platform. This includes making unsupported claims about the virus's origins or posts that downplay the seriousness of the pandemic, as well as baseless treatments or cures. |
| **Efficacy & safety of prevention & treatment measures** | Facebook, Instagram | More specifically, we remove false information about:<br>Guaranteed cures or prevention methods for COVID-19: Public health authorities, such as the WHO, say there is currently nothing that can guarantee recovery or guarantee the average person will not get COVID-19. We have also heard from public health authorities that if people thought there was a guaranteed cure or prevention for COVID-19, that could lead them to take incorrect safety measures, ignore appropriate health guidance, or even attempt harmful self-medication. This is why we don't allow false claims about how to cure or prevent COVID-19. This includes:<br>Claims that for the average person, something can guarantee prevention from getting COVID-19 or can guarantee recovery from COVID-19 before such a cure or prevention has been approved, including:<br>Consuming or inhaling specific items.<br>Medical or herbal remedies.<br>External remedies for the outer body or skin.<br>Ex: "Take Vitamin C – it cures COVID-19," "If you take this herbal remedy, you will not get COVID-19," "This topical cream will prevent you from contracting coronavirus." |
| | YouTube | Don't post content on YouTube if it includes any of the following:<br>Treatment misinformation:<br>Content that encourages the use of home remedies, prayer, or rituals in place of medical treatment such as consulting a doctor or going to the hospital.<br>Content that claims that there's a guaranteed cure for COVID-19.<br>Other content that discourages people from consulting a medical professional or seeking |

| | | |
|---|---|---|
| | | medical advice.<br>Prevention misinformation: Content that promotes prevention methods that contradict local health authorities or WHO.<br>Claims that there is a guaranteed prevention method for COVID-19. |
| | Twitter | We will label or remove false or misleading information about:<br>The safety or efficacy of treatments or preventative measures that are not approved by health authorities, or that are approved by health authorities but not safe to administer from home.<br>Under this guidance, we will require people to remove Tweets that include:<br>Misleading claims that unharmful but ineffective methods are cures or absolute treatments for COVID-19, such as "Coronavirus is vulnerable to UV radiation – walking outside in bright sunlight will prevent COVID-19."<br>Description of harmful treatments or preventative measures which are known to be ineffective or are being shared out of context to mislead people, such as "drinking bleach and ingesting colloidal silver will cure COVID-19." |
| | LinkedIn | We've always prohibited false and misleading content, but we recently updated our Professional Community Policies to be clear that information contradicting guidance from leading global health organizations and public health authorities is also not allowed on the platform. This includes making unsupported claims about the virus's origins or posts that downplay the seriousness of the pandemic, as well as baseless treatments or cures. |
| **COVID-19 vaccines** | Facebook, Instagram | Stringent Regulatory Authorities (SRAs) have issued emergency use authorization for several COVID-19 vaccines, so in addition to false claims about face masks, social distancing and testing, we do not allow false claims about the vaccines or vaccination programs that public health experts have advised us could lead to COVID-19 vaccine rejection. This includes false claims about the safety, efficacy, ingredients, development, existence, or conspiracies related to the vaccine or vaccination program. |
| | YouTube | Don't post content on YouTube if it includes any of the following:<br>Claims about COVID-19 vaccinations that contradict expert consensus from local health authorities or WHO. |
| | Twitter | We will label or remove false or misleading information about:<br>Vaccines and vaccination programs which suggest that COVID-19 vaccinations are part of a deliberate or intentional attempt to cause harm or control populations. |
| | LinkedIn | We're also continuing to keep our members safe and informed when it comes to trusted sources of vaccine news and information, and we are actively working to remove any misinformation about vaccines from our platform. |
| | TikTok | TikTok's Community Guidelines prohibit content that's false or misleading, including misinformation related to COVID-19 and vaccines and anti-vaccine disinformation more broadly. |
| **Restrictions & health advisories** | Facebook, Instagram | More specifically, we remove false information about:<br>Discouraging good health practices: There are a number of good health practices public health authorities advise people take to protect themselves from getting or spreading COVID-19. This includes wearing a face mask, social distancing, getting tested for COVID-19 and, more recently, getting vaccinated against COVID-19.<br>As more information becomes available about COVID-19 vaccines, we will continue to iterate on how we apply this policy. This includes:<br>Claims about wearing a face mask, including:<br>Claims that wearing a face mask does not help prevent the spread of COVID-19.<br>Claims that face masks include or are connected to 5G technology.<br>Claims that wearing a face mask can make the wearer sick.<br>Claims that public health authorities do not recommend that healthy people wear masks.<br>Claims that social/physical distancing does not help prevent the spread of COVID-19. |
| | YouTube | Don't post content on YouTube if it includes any of the following:<br>Social distancing and self isolation misinformation: Content that disputes the efficacy of local health authorities' or WHO's guidance on physical distancing or self-isolation measures to reduce transmission of COVID-19. |
| | Twitter | We will label or remove false or misleading information about:<br>Personal protective equipment (PPE) such as claims about the efficacy and safety of face |

| | | |
|---|---|---|
| | | masks to reduce viral spread.<br>Preventative measures such as hand-washing, proper hygiene or sanitation methods, or social distancing. |
| **Misrepresenting data** | Facebook, Instagram | We remove content that can contribute to physical harm by inaccurately representing the access to or availability of public health infrastructure. |
| | YouTube | Here are some examples of content that's not allowed on YouTube:<br>Claims that there have not been cases or deaths in countries where cases or deaths have been confirmed by local health authorities or the WHO. |
| | Twitter | We will label or remove false or misleading information about:<br>The prevalence of the virus or the disease, such as information pertaining to test results, hospitalizations, or mortality rates.<br>The capacity of the public health system to cope with the crisis, for example false information about the availability of PPE, ventilators, or doctors, or about hospital capacity.<br>Research findings (such as misrepresentations of or unsubstantiated conclusions about statistical data) used to advance a specific narrative that diminishes the significance of the disease. |

***Table 3.*** *Coded excerpts of criteria used by social media platforms to take action against misleading COVID-19 content.*

| | | |
|---|---|---|
| **Claims debunked by public health authorities** | Facebook, Instagram | We also remove false claims or conspiracy theories that have been flagged by leading global health organizations and local health authorities as having the potential to cause harm to people who believe them. |
| | YouTube | YouTube doesn't allow content that spreads medical misinformation that contradicts local health authorities' or the World Health Organization's (WHO) medical information about COVID-19. |
| | Twitter | Under this policy, we consider claims to be false or misleading if (1) they have been confirmed to be false by subject-matter experts, such as public health authorities. |
| | LinkedIn | We've always prohibited false and misleading content, but we recently updated our Professional Community Policies to be clear that information contradicting guidance from leading global health organizations and public health authorities is also not allowed on the platform. |
| **Potential of the content to lead to significant ham** | Facebook | Under our Community Standards, we remove misinformation when public health authorities conclude that the information is false and likely to contribute to imminent violence or physical harm. Since COVID-19 was declared a Public Health Emergency of International Concern (PHEIC) in January 2020, we have applied this policy to content containing claims related to COVID-19 that, according to public health authorities, are (a) false, and (b) likely to contribute to imminent physical harm (of imminent physical harm examples include: increasing the likelihood of exposure to or transmission of the virus, or having adverse effects on the public health system's ability to cope with the pandemic). |
| | Instagram | We also remove false claims or conspiracy theories that have been flagged by leading global health organizations and local health authorities as having the potential to cause harm to people who believe them.<br>We've connected over 2 billion people from 189 countries to reliable information about the coronavirus through our COVID-19 Information Center and informational messages, and we've removed more than 12 million pieces of content on Facebook and Instagram containing misinformation that could lead to imminent physical harm. |
| | YouTube | YouTube doesn't allow content about COVID-19 that poses a serious risk of egregious harm. |
| | Twitter | Content that is demonstrably false or misleading and may lead to significant risk of harm (such as increased exposure to the virus, or adverse effects on public health systems) may not be shared on Twitter. |
| | Snapchat | Our guidelines prohibit Snapchatters and our partners from sharing content that deceives or deliberately spreads false information that causes harm, and we do not offer an open news feed where unvetted publishers or individuals have an opportunity to broadcast misinformation. |

| | | |
|---|---|---|
| | Reddit | The situation on the ground is constantly changing and so we are trying to strike a balance of acting quickly on claims that might cause or encourage violence or physical harm (such as advice to drink bleach, or calls to vandalize phone towers). |
| | TikTok | Our Community Guidelines prohibit misinformation that could cause harm to our community or the larger public, including content that misleads people about elections or other civic processes, content distributed by disinformation campaigns, and health misinformation. |
| **Inauthentic content** | Facebook, Instagram | Over the past three years, we've removed over 100 networks of coordinated inauthentic behavior (CIB) from our platform and keep the public informed about our efforts through our monthly CIB reports. |
| | YouTube | Our guidelines against deceptive practices include tough policies against users who misrepresent themselves or who engage in other deceptive practices. This includes deceptive use of manipulated media (e.g., 'deep fakes') which may pose serious risks of harm. |
| | Twitter | You can't create fake accounts which misrepresent their affiliation, or share content that falsely represents its affiliation to a medical practitioner, public health official or agency, research institution, or that falsely suggests expertise on COVID-19 issues. |
| | Snapchat | We regularly review and update our policies as new forms of misinformation become more prevalent: for example, ahead of the 2020 election, we updated our guidelines to make clear that manipulated media intended to mislead -- or deepfakes -- were prohibited. |
| | Reddit | Our site integrity team is using their existing tools and processes to investigate claims and signs of coordinated attempts to spread COVID-19 misinformation on Reddit. |
| | TikTok | We're adding a policy which prohibits synthetic or manipulated content that misleads users by distorting the truth of events in a way that could cause harm. Our intent is to protect users from things like shallow or deep fakes, so while this kind of content was broadly covered by our guidelines already, this update makes the policy clearer for our users. |
| | WhatsApp | We've also set a limit on the number of times messages can be forwarded on WhatsApp to reduce the spread of viral messages, and we use advanced machine learning to identify and ban accounts engaged in mass messaging. |
| **Other** | Facebook, Instagram | As people around the world confront this unprecedented public health emergency, we want to make sure that our Community Standards protect people from harmful content and new types of abuse related to COVID-19. We're working to remove content that has the potential to contribute to real-world harm, including through our policies prohibiting the coordination of harm, the sale of medical masks and related goods, hate speech, bullying and harassment, and misinformation that contributes to the risk of imminent violence or physical harm. |
| | YouTube | Content that comes close to — but doesn't quite cross the line of — violating our Community Guidelines is a fraction of 1% of what's watched on YouTube in the U.S. Our recommendations systems do not recommend such content on YouTube, thereby helping limit the spread of borderline content or videos that could misinform users in harmful ways. |
| | Twitter | In order for content related to COVID-19 to be labeled or removed under this policy, it must:<br>Advance a claim of fact, expressed in definitive terms.<br>For a Tweet to qualify as a misleading claim, it must be an assertion of fact (not an opinion), expressed definitively, and intended to influence others' behavior.<br>Going forward and specific to COVID-19, unverified claims that have the potential to incite people to action, could lead to the destruction or damage of critical infrastructure, or cause widespread panic/social unrest may be considered a violation of our policies. Examples include, "The National Guard just announced that no more shipments of food will be arriving for two months — run to the grocery store ASAP and buy everything" or "5G causes coronavirus — go destroy the cell towers in your neighborhood!" |

***Table 4.*** *Coded excerpts of remedies to address COVID-19 misinformation across leading social media and messaging platforms.*

| | | |
|---|---|---|
| **Labels, warnings, notifications & links** | Facebook | For the duration of the COVID public health emergency, we also remove certain COVID-19 misinformation. In our third-party fact-checking program, fact-checkers rate and review all types of content, and we add a warning label with more information and reduce its distribution. |
| | Instagram | Our global network of third-party fact-checkers are continuing their work reviewing content and debunking false claims that are spreading related to the coronavirus. When they rate information as false, we limit its spread on Facebook and Instagram and show people accurate information from these partners. We also send notifications to people who already shared or are trying to share this content to alert them that it's been fact-checked. |
| | Twitter | Starting today, we're introducing new labels and warning messages that will provide additional context and information on some Tweets containing disputed or misleading information related to COVID-19. |
| | Reddit | A quarantine will remove the community from search results, warn the user that it may contain misinformation, and require an explicit opt-in. |
| | TikTok | For TikTok users who choose to explore hashtags related to coronavirus, we surface an in-app notice that provides direct access to WHO's website and local public health agencies while also reminding users to report content that violates our Community Guidelines. |
| | Messenger, WhatsApp | On WhatsApp and Messenger: We've built clear labels that show people when they have received a forwarded message, or chain message, so they know when they are receiving something that was not written by their immediate contacts. |
| **Decreasing visibility & spread** | Facebook | Pages, Groups, profiles, and Instagram accounts that repeatedly post misinformation related to COVID-19, vaccines, and health may face restrictions, including (but not limited to) reduced distribution, removal from recommendations, or removal from our site. |
| | Instagram | As part of our efforts to improve the quality of health and vaccine content that people encounter during the COVID-19 pandemic, and consistent with the advice of independent health experts, we are also taking additional steps to reduce the distribution of certain other content about vaccines that does not otherwise violate our policies listed above, and remove certain Pages, Groups, and Instagram accounts that have shared content that violates our COVID-19 and vaccine policies and are dedicated to spreading vaccine discouraging information on platform. Specifically, we are taking additional steps to limit visibility of this content on our recommendations surfaces. |
| | YouTube | Over the past few years, we've accelerated our efforts to protect the YouTube community from harmful content. This is also how we are approaching COVID-19-related content on YouTube. We raise authoritative voices, remove misinformation and reduce the spread of borderline content so that our community can connect with timely and helpful information at this critical time. |
| | Twitter | In addition, we're halting any auto-suggest results that are likely to direct individuals to noncredible content on Twitter. |
| | Reddit | A quarantine will remove the community from search results, warn the user that it may contain misinformation, and require an explicit opt-in. |
| | TikTok | We take multiple approaches to make anti-vaccine and COVID-19 misinformation harder to find. In addition to removing content, we redirect searches associated with vaccine or COVID-19 disinformation to our Community Guidelines and do not autocomplete anti-vaccine hashtags in search. |
| | Messenger | As a part of our ongoing efforts to provide people with a safer, more private messaging experience, today we're introducing a forwarding limit on Messenger, so messages can only be forwarded to five people or groups at a time. Limiting forwarding is an effective way to slow the spread of viral misinformation and harmful content that has the potential to cause real world harm. |
| | WhatsApp | We've also set a limit on the number of times messages can be forwarded on WhatsApp to reduce the spread of viral messages, and we use advanced machine learning to identify and ban accounts engaged in mass messaging. |

| | | |
|---|---|---|
| **Content removal** | Facebook, Instagram | We will, however, remove certain COVID-19 misinformation that has been previously debunked by multiple independent fact-checkers. |
| | YouTube | If your content violates this policy, we'll remove the content and send you an email to let you know. |
| | Twitter | COVID-19 related content that meet all three of the criteria defined above—i.e. that are claims of fact, demonstrably false or misleading, and likely to cause harm—may not be shared on Twitter and are subject to removal. |
| | LinkedIn | We're also continuing to keep our members safe and informed when it comes to trusted sources of vaccine news and information, and we are actively working to remove any misinformation about vaccines from our platform. |
| | Snapchat | Our approach to enforcing against content that includes false information is straightforward -- we don't label it, we completely remove it. When we find content that violates our guidelines, our policy is to simply take it down, which immediately reduces the risk of it being shared more widely. |
| | Reddit | We've already seen many of you stepping up to set up automod rules to remove the most obvious pieces of misinformation. |
| | TikTok | We take multiple approaches to make anti-vaccine and COVID-19 misinformation harder to find. In addition to removing content, we redirect searches associated with vaccine or COVID-19 disinformation to our Community Guidelines and do not autocomplete anti-vaccine hashtags in search. |
| **Account suspension/ban** | Facebook, Instagram | Pages, Groups, profiles, and Instagram accounts that repeatedly post misinformation related to COVID-19, vaccines, and health may face restrictions, including (but not limited to) reduced distribution, removal from recommendations, or removal from our site |
| | YouTube | If this is your first time violating our Community Guidelines, you'll get a warning with no penalty to your channel. If it's not, we'll issue a strike against your channel. If you get 3 strikes, your channel will be terminated. You can learn more about our strikes system here. |
| | Twitter | For severe or repeated violations of this policy, accounts will be permanently suspended. |
| | Reddit | We are taking several actions:<br>Ban r/NoNewNormal immediately for breaking our rules against brigading. |
| | TikTok | Our Community Guidelines and Terms of Service apply to everyone who uses TikTok and all content they post. We use a mix of technology and human moderation to enforce these policies, including by removing content, banning accounts, and making it more difficult to find harmful content, like misinformation and conspiracy theories, in recommendations or search. |
| | WhatsApp | We've also set a limit on the number of times messages can be forwarded on WhatsApp to reduce the spread of viral messages, and we use advanced machine learning to identify and ban accounts engaged in mass messaging. |
| **Ad restrictions** | Facebook | Under our Regulated Goods policy, we've taken steps to protect against exploitation of this crisis for financial gain and prohibit the below content when we have additional information and/or context to identify it: Makes mention of medical products and COVID-19 and indicates a sense of urgency or claims that prevention is guaranteed. |
| | Instagram | To prevent people from exploiting this public health emergency we've already put several new policies into effect. We prohibited misleading ads for products that refer to COVID-19 in ways intended to create urgency, guarantee cures or prevent people from contracting it. |
| | YouTube | All monetizing content is subject to our Ad Friendly Guidelines and Community Guidelines. If your content violates these policies, it will be removed or receive limited or no ads. For specific examples of COVID-19 related content that isn't eligible for monetization, check out this Help Center article. |
| | Twitter | The following restrictions apply to these use cases:<br>Distasteful references to COVID-19 (or variations) are prohibited.<br>Content may not be sensational or likely to incite panic.<br>Prices of products related to COVID-19 may not be inflated.<br>The promotion of certain products related to COVID-19 may be prohibited. |

| | | |
|---|---|---|
| | LinkedIn | Microsoft's Sensitive Advertising policy and LinkedIn's Ads Policies prohibit ads that capitalize on the pandemic and company pages that improperly sell medical supplies and solutions. These policies allow Microsoft and LinkedIn to remove or limit advertising and company pages in response to a sensitive tragedy, disaster, death or high-profile news event, and are being applied to block ads related directly to COVID-19. Any advertising that exploits the coronavirus crisis for commercial gain, spreads misinformation or might pose a danger to users' safety is prohibited. |
| | Snapchat | **We use human review to fact check all political and advocacy ads.** As with all content on Snapchat, we prohibit false information and deceptive practices in our advertising. All political ads, including election-related ads, issue advocacy ads, and issue ads, must include a transparent "paid for" message that discloses the sponsoring organization. We use human review to fact check all political ads, and provide information about all ads that pass our review in our Political Ads library. |
| | TikTok | We also do not allow paid advertising that advocates against vaccinations, though PSAs or calls to action related to COVID-19 vaccines are accepted on a case-by-case basis if they're in the interest of public health and safety. |

***Table 5.*** *Coded excerpts of remedies to promote access to evidence-based COVID-19 information across leading social media platforms.*

| | | |
|---|---|---|
| **Information curation** | Facebook | Facebook COVID-19 Information Center |
| | Instagram | Today, we're bringing the COVID-19 Information Center to Instagram all around the world. This portal, which we launched in the Facebook app last March, helps people discover the latest information about the virus from local health ministries and the World Health Organization. |
| | YouTube | A COVID-19 news shelf may now show on the YouTube homepage. The shelf includes news videos about COVID-19 from authoritative news publishers and local health authorities on our platform. The content in this shelf is populated algorithmically, using hundreds of signals, including relevance to COVID-19, how up-to-date it is, and region. |
| | Twitter | We've added a new tab in Explore so it's easier to find the latest information on COVID-19. The tab will include curated pages highlighting the latest news such as public service announcements, Tweets from public health experts and journalists, as well as stories about how people are coping and helping each other. |
| | LinkedIn | Since early March, we've delivered news and perspectives about the coronavirus from official and trusted sources, all curated by our team of 65+ LinkedIn Editors. We've created a "Special Report: Coronavirus" box on the top right of our homepage and ensured that when members search for terms or hashtags related to the virus they see this coverage first. Through our Daily Rundown feature, which reaches 46 million people in 96 countries, we deliver timely and relevant updates to our members. |
| | Snapchat | The WHO and CDC publish regular updates for Snapchatters from their Official Accounts and we've worked with the WHO to develop custom content to answer questions from our community. |
| | TikTok | TikTok works with public health experts to make authoritative information about COVID-19 and vaccines available directly in our app. In our COVID-19 information hub, our community can find answers to common questions about the virus and vaccines from the World Health Organization (WHO) and the CDC as well as tips on staying safe. |
| | Messenger | Today the World Health Organization (WHO) launched an interactive experience on Messenger to provide accurate and timely information about the coronavirus outbreak. People will now be able to message the WHO with questions about COVID-19 and get quick answers for free. |
| | WhatsApp | Today we launched the World Health Organization's Health Alert on WhatsApp. The WHO Health Alert is free to use and will answer common questions about COVID-19. It provides timely, reliable information about how to prevent the spread of the coronavirus as well as travel advice, coronavirus myth debunking and more. |
| **Health promotion campaigns** | Facebook | With the rise in COVID-19 cases in the US and in many other parts of the world, we are expanding our alerts reminding people to wear face coverings internationally as recommended by health authorities. These alerts have been running at the top of Facebook and Instagram in the US since early July. |

| | | |
|---|---|---|
| | Instagram | We're also launching new stickers to help people share accurate COVID-19 information in Stories. These new features include reminders to wash your hands, distance yourself from others and more. These will be available in the camera in the coming days. |
| | YouTube | We are working with a wide range of partners who have experience and expertise in public health communication with key at-risk groups, to help create, amplify and promote their campaigns on YouTube, including the Kaiser Family Foundation's Greater Than COVID video campaign, featuring their series for Black America and two upcoming series that seek to reach both the Latinx community and low-income rural communities across America; and the Black Coalition Against COVID-19 and Black Doctor Org video series to help answer top questions from the Black community about COVID and vaccines. |
| | Twitter | In January, in partnership with Team Halo, UNICEF, NHS, and the Vaccine Confidence Project, we activated an emoji hashtag #vaccinated to show support for vaccination. This builds upon our earlier efforts to encourage people to #StayHome, #WashHands, and #WearAMask. |
| | Snapchat | Bitmojis: Add Bitmojis to your Snaps to share tips and spread awareness. Try searching 'Wash your hands,' 'Stay home,' 'Don't touch your face,' and 'Social distancing' when selecting a Bitmoji! |
| | TikTok | TikTok supported the Safe Hands Challenge, a campaign launched by the WHO to promote hand washing. The hashtag has 5.4 billion views with participants like Jimmy Fallon, Gloria Gaynor, and Mariah Carey. |
| **Labels, banners & links** | Facebook, Instagram | On Facebook and Instagram: In January, we started showing educational pop-ups connecting people to information from the WHO, the CDC and regional health authorities toward the top of News Feed in countries with reported person-to-person transmissions and in all countries when people search for COVID-19 related information. |
| | YouTube | As a continuation of our efforts to combat COVID-19 related misinformation, we're updating our COVID-19 information panels to include links to COVID-19 vaccine info. The updated panels may show in search results and on watch pages related to COVID-19 or COVID-19 vaccine info. The updated panels are intended to help users find third-party authoritative COVID-19 vaccine info and are not a judgment on the accuracy of any video. |
| | Twitter | Since January 2020, we have had a dedicated COVID-19 search prompt feature in place within the product. This means when someone searches for COVID-19, they are met with credible, authoritative content at the very top of their search experience. This has been expanded to over 80 countries worldwide and is currently available in 29 languages. In some countries the prompts now also include an additional button which links to COVID-19 vaccine specific information. |
| | LinkedIn | We're taking new steps across our services, including Bing, LinkedIn, Microsoft News and Microsoft Advertising, to include curated resources on Microsoft News and LinkedIn that link to official guidance from organizations such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC). |
| | Snapchat | We launched creative tools to help Snapchatters share expert-approved best practices with their friends and family members, including a worldwide Filter with advice to our community on how to stay safe. This information is sourced from the World Health Organization, and links to its website for more info. |
| | Reddit | We have labels on coronavirus-related videos which point users to trusted information, with resources directly in the app as well as in the dedicated COVID-19 section of our Safety Center. |
| | TikTok | The informational hub can be accessed from the Discover page, search, and banners on videos related to COVID-19 and vaccines. |
| **Increasing visibility of authoritative content** | Facebook | Building on our goal to promote authoritative information about COVID-19 vaccines, we have implemented several temporary measures to further limit the spread of potentially harmful COVID-19 and vaccine information during the pandemic. Some of these measures include: continuing to further elevate information from authoritative sources when people seek information about COVID-19 or vaccines. |

| | | |
|---|---|---|
| | Instagram | To help people get relevant and up to date resources, we will start showing more information from @WHO and local health ministries at the top of Instagram's feed in some countries. |
| | YouTube | This is also how we are approaching COVID-19-related content on YouTube. We raise authoritative voices, remove misinformation and reduce the spread of borderline content so that our community can connect with timely and helpful information at this critical time. |
| | Twitter | Since January 2020, we have had a dedicated COVID-19 search prompt feature in place within the product. This means when someone searches for COVID-19, they are met with credible, authoritative content at the very top of their search experience. |
| | LinkedIn | And when users search for coronavirus-related terms or hashtags, they'll see trusted information modules at the top of the results page. |
| **Q&As with experts** | Facebook | Mark Zuckerberg is live with Dr. Anthony Fauci, America's top infectious disease expert, to discuss progress toward a COVID-19 vaccine and how we can slow the spread of the virus this holiday season. |
| | YouTube | We're building on the success of conversations like the ones between Dr. Fauci and Monica, CDC officer Tia Rogers and Asia Jackson, and Andy Slavitt and Jim Gaffigan, to connect with more audiences, ranging from rural and farming communities to family vloggers. |
| | Twitter | We continue to host a weekly live Q&A event page for the WHO at #AskWHO. |
| | LinkedIn | The WHO is updating daily with live streams of their media briefings, tips to stay safe and healthy during the pandemic, and hosting real-time Q&As with experts, which is generating some of the highest views on LinkedIn Live. |
| | Snapchat | Our own news team is also regularly producing coverage and continuously updating Discover with tips and information about COVID-19, including Q&As with medical experts. |
| | Reddit | We're also continuing to curate an expert AMA series so we can give you direct access to scientific and medical professionals and relevant public officials. |
| | TikTok | We also hosted a series of live streams led by the World Health Organization where experts from WHO shared information on protective measures and took live questions from our users. |
| **Ad credits to public health partners** | Facebook, Instagram | On Facebook and Instagram: We're also giving the WHO as many free ads as they need and millions in ad credits to other health authorities so they can reach people with timely messages. |
| | YouTube | We've also donated ad inventory to governments and NGOs to help give their public health messages about COVID-19 more visibility on YouTube. |
| | Twitter | In addition, we have donated premium advertising products, including Promoted Trend and First View products, to elevate critical public health information such as @FEMA's message about the agency's vaccination efforts and emergency relief locations during winter storms. |
| | LinkedIn | We are also providing free ads to organizations that will disseminate critical information on Covid-19 vaccines such as the UN Verified Initiative, World Health Organization and The Ad Council. |
| | TikTok | Around the world, health authorities are working to inform the public as quickly as possible on a range of issues, including the importance of social distancing or proper hygiene. To facilitate that education, we are providing $25M in prominent in-feed ad space for NGOs, trusted health sources, and local authorities, enabling them to share important messages with millions of people and meaningfully engage the TikTok community. |